Detection of Children's Paralinguistic Events in Interaction with Caregivers

[1]Hrishikesh Rao, [1]Jonathan C. Kim, [1]Mark A. Clements, [2]Agata Rozga, and [3]Daniel S. Messinger

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA
[3]Department of Psychology, University of Miami, Coral Gables, FL, USA
{hrishikesh, jon.kim, clements, agata}@gatech.edu, dmessinger@miami.edu

Abstract

Paralinguistic cues in children's speech convey the child's affective state and can serve as important markers for the early detection of autism spectrum disorder (ASD). In this paper, we detect paralinguistic events, such as laughter and fussing/crying, along with toddlers' speech from the Multi-modal Dyadic Behavior Dataset (MMDB). We use both spectral and prosodic acoustic features selected using a combination of filter and wrapper-based methods. The classification accuracy using a support vector machine with a linear kernel for detecting laughter in children's speech was 77.87% and that for fussing/crying was 79.37%. A tertiary classification scheme for detecting laughter, fussing/crying, and speech yielded an accuracy of 69.73%. To test for the generalization of the approach for detecting fussing/crying, we used recordings from the Strange Situation protocol, which is used to observe attachment behavior between an infant and a parent. Using a cross-corpus testing set for detecting fussing/crying, we obtained a detection accuracy of 71.6%. These results indicate that the selected acoustic features are capable of discriminating children's laughter, fussing/crying, and speech and the algorithms generalize well to a dataset consisting of paralinguistic cues of a different age group, infants (12 - 18 months of age), gathered in a different context.

Index Terms: laughter, crying, children, affect, acoustic analysis.

10.21437/Interspeech.2014-309

1. Introduction

Paralinguistic cues, such as laughter and crying, play an important role in children's early communication, and these cues are useful in conveying the affective state of the speaker. These cues have been found to be important markers in the very early detection of autism spectrum disorder (ASD) [1, 2], and the diarization of such events in extended recordings can be a useful aid in the diagnosis of developmental disorders [3, 4]. It can also be used to analyze children's communicative behaviors in social interactions with their caregivers. The main focus of our work is to detect laughter and fussing/crying in toddlers' speech using acoustic features. Laughter is primarily used to express positive affect and has been found to usually follow a state of anticipatory arousal, especially tickling [5]. Fussing/Crying could indicate that the child is upset or disinterested in the task being initiated by the caregiver in a dyadic setting.

We have used two datasets, the Multi-modal Dyadic Behavior (MMDB) dataset and the Strange Situation corpus, for the purpose of developing detectors for laughter, fussing/crying, and child's speech consisting of verbalizations and vocalizations. The spectral and prosodic features were extracted using openSMILE [6], Praat [7], and VoiceSauce [8]. We used a combination of wrapper and filter-based feature selection approaches to reduce the dimensionality of the feature set. The results of a 10-fold cross-validation show that our method of selecting features and trained models is capable of predicting laughter and fussing/crying in children's speech robustly. A test set using 11 MMDB sessions resulted in better than chance results of **77.87%** for predicting laughter, **79.37%** for predicting fussing/crying, and **69.73%** for the tertiary classification task of predicting speech, laughter, and fussing/crying. We developed a cross-corpus testing set of the MMDB and Strange Situation datasets and obtained an accuracy of **71.6%**. The importance of the results can be gauged from the fact that the datasets are disparate in not only the age range but also the type of fussing/crying samples.

The paper describes the corpora used in our analysis in Section 2. The acoustic features extracted are described in Section 3 and Section 4 discusses the feature selection techniques employed. The interpretation of the features selected is summarized in Section 5 and finally, the results are discussed in Section 6.

2. Corpora

The datasets that have been employed in this study are the Multi-modal Dyadic Behavior (MMDB) dataset [9] recorded at the Child Study Lab (CSL) at the Georgia Institute of Technology, Atlanta, GA, USA and a set of 10 practice Strange Situations that had been conducted in multiple laboratories and were nationally distributed by researchers at the University of Minnesota, Minneapolis, MN, USA (http://attachmenttraining.com).

2.1. Multi-modal Dyadic Behavior Dataset

The MMDB dataset consists of a brief (3-5 minutes) semi-structured interaction between a child (15-30 months of age) and an adult while seated at a table across from each other. The protocol is designed to elicit the child's social attention, back-and-forth interaction, and non-verbal communication. The protocol consists of five activities: greeting the child, initiating a game of rolling a ball back and forth, bringing a book and inviting the child to look through it, placing the book on the head and pretending it to be a hat, and engaging the child in a tickling game. For our analysis, we focused on the child's verbal behavior for detecting instances of laughter, fussing/crying, and speech which were annotated by two research assistants in the CSL. We have selected 35 sessions, which constitutes the training data, for detecting the child's paralinguistic events (laughter and fussing/crying) and the speech. The test set consists of 11 sessions. The ages of the participants ranged from 15-30

months with a mean of 21.65 months and a standard deviation of 4.84 months. The number of samples along with the mean and standard deviation of the duration of the samples of laughter, fussing/crying, and speech of the training and test sets are shown in Table 1. Owing to the large number of samples of children's speech and to prevent overfitting of the training data, the speech class was balanced by randomly selecting 58 samples.

Table 1: *Number of training and testing examples of MMDB dataset for speech, laughter, and fussing/crying along with the mean and standard deviation of duration of the samples.*

Dataset	Type of Vocalization	Number of samples (N)	Duration (s) (mean$\pm$standard deviation)
Training Set	Speech (before balancing)	501	1$\pm$0.87
	Speech (after balancing)	58	1.14$\pm$0.66
	Laughter	54	1.31$\pm$1.28
	Fussing/Crying	62	2.65$\pm$4.21
Testing Set	Speech	122	1.23$\pm$0.92
	Laughter	35	1.12$\pm$0.90
	Fussing/Crying	30	1.68$\pm$0.83

2.2. Strange Situation Dataset

Recordings were made during the Strange Situation procedure [10]. The procedure consists of eight 3-minute episodes including two separations from the mother, each followed by a reunion [11]. The episodes are arranged in a manner to create a series of stressful situations for the infant. The goal is to evaluate how the child reacts to being reunited with the mother, specifically, whether he/she approaches her, is soothed by the contact, and returns to play. The detection of crying is an important behavior considered in the scoring of this assessment. In this dataset, only the fussing/crying events were annotated (N=62). The mean duration of the samples was 4.35 seconds and the standard deviation was 4.62 seconds.

The type of fussing/crying differs in both the corpora. The subjects in the MMDB dataset usually whimper to indicate discomfort with the activities while in the Strange Situation recordings the subjects cry when they are separated from the mother.

3. Feature Extraction

The acoustic features were extracted using the open-source audio feature extraction tool, openSMILE [6]. There were 57 low-level descriptors (LLD), shown in Table 2 extracted using a 30 ms Hamming window with 10 ms overlap. The delta and delta-delta measure for each LLD was also computed and the number of LLDs was 171. There were 39 statistical measures, shown in Table 3, computed from the LLDs for each sample. The dimensionality of the feature set using openSMILE was 6669.

The formant-based features were extracted using Praat [7] and the cepstral peak prominence (CPP) was extracted using VoiceSauce [8]. The first four formant frequencies, resonances in the vocal tract [12], and their respective bandwidths were extracted. The delta and delta-delta for the formant-based frequencies were also extracted. The CPP is a measure of breathiness in speech and is computed by measuring the difference between the peak of the cepstrum and a linear regression line fitted to the cepstrum [13]. It gives the measure of the periodicity of the signal. These features are shown in Table 4. The total number of low-level descriptors for formant and CPP-based features was 25. The statistical measures, shown in Table 5, were computed for these features and the dimensionality of the formant-based and CPP features was 350.

Table 2: *Spectral and prosodic acoustic features extracted using openSMILE.*

Feature	Number of low-level descriptors
Log-energy	3
Magnitude of Mel-Spectrum	78
Mel-frequency Cepstral Coefficients	39
Pitch	3
Pitch envelope	3
Probability of voicing	3
Magnitude in frequency band ($0-250Hz$, $250-650Hz$, $0-650Hz$, $1000-4000Hz$, and $3010-9123Hz$)	16
Spectral Rolloff (25^{th} ,50^{th}, 75^{th}, and 90^{th} percentile)	12
Spectral Flux	3
Spectral Position (Centroid, Maximum, and Minimum)	3
Zero-Crossing Rate	3

Table 3: *Statistical measures evaluated for openSMILE features.*

Statistical Measure
Max./Min. value and respective relative position within input, range, arithmetic mean,3 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, centroid, variance, number of non-zero elements, quadratic, geometric, absolute mean, arithmetic mean of contour and non-zero elements of contour, 95^{th} and 98^{th} percentiles, number of peaks, mean distance from peak, mean peak amplitude, quartile 1 - 3, and 3 inter-quartile ranges.

Table 4: *Formant-based and cepstral peak prominence features.*

Feature	Number of low-level descriptors
Formant frequency	12
Formant bandwidth	12
Cepstral peak prominence	1

Table 5: *Statistical measures evaluated for formant-based and cepstral peak prominence features.*

Statistical Measure
Arithmetic mean, median, mode, standard deviation, maximum and minimum values, flatness, skewness, kurtosis, 25^{th} quartile, 75^{th} quartile, inter-quartile ranges, 1^{st} percentile, 99^{th} percentile

4. Feature Selection

We have used a two-pronged approach of using a filter and wrapper-based approach for feature selection. This incorporates the advantages of evaluating the intrinsic properties of the dataset using the filter-based method and the ability to generalize well by avoiding overfitting using the wrapper-based method. There is an added benefit of reduction in computation time by selecting the k top features using the filter-based method and performing a wrapper-based feature selection on the reduced dimensionality feature set. The wrapper-based approach employs the correlation-based (CFS) and information gain ratio (IGR) feature selection techniques.

4.1. Correlation-based Feature Selection

The CFS [14] method selects features that are highly correlated with the class and uncorrelated with each other. For a subset of features S which contains k features and c classes, let r_{cf} be the mean feature-class correlation and r_{ff} be the mean feature-feature correlation, then the heuristic merit M_s is computed as shown in (1),

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \quad (1)$$

The CFS method, in our paper, evaluates the correlation between a feature (k=1) and the class. The correlation between the feature and the class is computed using the Pearson correlation coefficient.

4.2. Information Gain Ratio Feature Selection

The information gain ratio (IGR) [15] is the information gain normalized by the intrinsic information of the feature. The information gain measures the number of bits of information obtained for class prediction by knowing the presence or absence of a sample point in the classes [16].

Let $\{w_i\}_{i=1}^{M}$ be the set of classes and for any attribute, $\{X_j\}_{j=1}^{N}$, which has been discretized to N levels, the information gain of the attribute is given in (2).

$$IG(w_i, X_j) = H(w_i) - H(w_i|X_j), \quad (2)$$

where $H(w_i)$ is the entropy of the class w_i and $H(w_i|X_j)$ is the conditional entropy of the class w_i given the discretized attribute X_j. The problem with the information gain criterion is that it favors features with a large number of values [15] and can cause overfitting.

The intrinsic information of the feature is computed by measuring the entropy in the class as shown in (3).

$$IV(X_j) = H(X_j), \quad (3)$$

where $H(X_j)$ is the entropy of the feature.

Features with high intrinsic value are considered to be less useful in discriminating between classes. The IGR, shown in (4), reduces the bias towards multi-valued features.

$$GR(w_i, X_j) = \frac{H(w_i) - H(w_i|X_j)}{H(X_j)}, \quad (4)$$

The wrapper method which gives the highest accuracy for the openSMILE features, since they form the majority of the features in the set, using a 10-fold cross-validation using a support vector machine (SVM) with sequential minimization optimization (SMO) for the binary classification tasks and a multi-class one-class-against-all SVM for the tertiary classification task, was selected as shown in Table 6 and used as the intermediate feature set for the filter-based method. For the binary classification tasks, the threshold for ranking and selecting the openSMILE features was 100 and for the tertiary classification the threshold was 200. The higher threshold for the tertiary scheme would enable the multi-class one-class-against-all classifier to discriminate between one class and the other classes which are treated as a singular class. For the formant and CPP-based features, the threshold was 50 for the three schemes.

For the binary selection tasks, the CFS method is used to select the top 100 openSMILE and 50 formant and CPP-based features. For the tertiary classification, IGR is employed to extract the top 200 openSMILE features and 50 formant and CPP-based features. It is of interest to note, from Table 6 that the results are better than chance for both the feature sets for all the classification tasks.

Table 6: *Results of 10-fold cross-validation using a support vector machine (SVM) with linear kernel for the wrapper-based feature selection methods for the openSMILE features along with results for formant and CPP-based features.*

Classification Task	Wrapper-based Feature Selection	Accuracy	
		openSMILE features	Formant and CPP-based features
Speech vs. Laughter	CFS	**81.3**%	79.5%
	IGR	75.9%	75.9%
Speech vs. Fussing/Crying	CFS	**83.3**%	67.5%
	IGR	78.3%	68.3%
Speech vs. Laughter vs. Fussing/Crying	CFS	68.4%	56.6%
	IGR	**70.1**%	60.1%

4.3. Sequential forward selection

The sequential forward selection (SFS) employs a support vector machine (SVM) with Sequential Minimization Optimization (SMO) and a linear kernel for the binary classification tasks. The tertiary classification scheme employs a Multi-class classifier using a one class-against-all SVM with SMO and a linear kernel. This method selects the feature which generates the highest accuracy in the feature set and iteratively adds features to the set until there is no more improvement in the accuracy. The methodology employed in this study is shown in Fig 1.

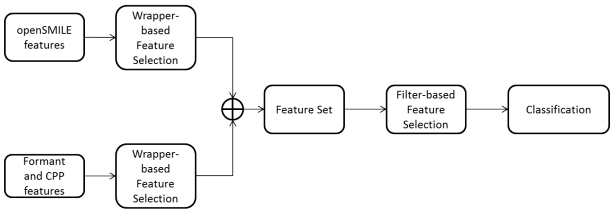

Figure 1: *Method for selection of features using wrapper and filter-based feature selection methods for classification.*

The purpose of the study is to understand which features are meaningful in discriminating between laughter, fussing/crying, and speech. As mentioned in Section 3, the two groups of features employed were the openSMILE and the formant and CPP-based features. These features were concatenated, after the wrapper-based feature selection, to form a 150-dimension feature set for the binary classification tasks and for the tertiary classification task, the dimensionality of the feature set was 250. The features were then processed to remove those with missing values and having less than 30% unique values. This was done to ensure that outliers did not affect the classification results. The SFS feature selection algorithm was used to further reduce the dimensionality of the feature set. The features and the number of statistical measures, selected using the wrapper and filter-based approaches, for the three classification schemes are shown in Table 7.

5. Feature Interpretation

The mel-frequency cepstral coefficients (MFCC) and the spectrum of the mel-spectrum constitute a major chunk of the features selected for the binary and tertiary classification tasks. The

Table 7: *Features selected for binary and tertiary classification tasks using combination of wrapper and filter-based features selection methods.*

Feature	Number of statistical measures		
	Speech vs. Laughter	**Speech vs. Fussing/Crying**	**Speech vs. Laughter vs. Fussing/Crying**
Mel-frequency cepstral coefficient	1	6	3
Magnitude of mel-cepstrum	2	2	1
Pitch	-	2	1
Probability of voicing	1	-	1
Log Energy	-	-	1
Cepstral Peak Prominence	2	-	-
Spectral Rolloff	1	-	-
Spectral Centroid	1	-	-
Fourth formant bandwidth	1	-	-

MFCCs, which mimic the human perception process, have been found to discriminate between adolescents' speech and laughter [16]. It has also been found to be useful in speaker identification of infants [17] and for pathological analysis using cries in infants [18]. The pitch-related features, probability of voicing, pitch, and cepstral peak prominence, have been found useful for all the classification tasks. These features are useful for discriminating between speech and laughter due to the consonant-vowel structure of laughter. Whining or fussing has been found to exhibit higher pitch and varied pitch contours [19] compared to adult-directed speech in children.

6. Results

For the purpose of classification, training models were developed using the three reduced feature sets from the MMDB dataset. The classifier used is an SVM with SMO with a linear kernel and the open-source classification tool, WEKA [20], was used for this purpose. A 10-fold cross-validation was performed on the three datasets and the results are shown in Table 8.

Table 8: *Classification results using a 10-fold cross-validation scheme with support vector machine (SVM) with a linear kernel.*

Classification Scheme	Average Recall	Average Accuracy
Speech vs. Laughter	**92.78%**	**92.85%**
Speech vs. Fussing/Crying	**88.49%**	**90.00%**
Speech vs. Fussing/Crying vs. Laughter	**76.43%**	**76.43%**

The results indicate that the laughter, fussing/crying, and speech can be discriminated robustly with the binary and tertiary classification schemes.

In order to test for the generalization of the results, a test set was devised consisting of 11 sessions selected randomly from the MMDB dataset. Seven sessions were used for the binary classification task of detecting laughter and the remaining 4 were used for detecting fussing/crying. The tertiary classification task used the combination of these. A grid search was performed by varying the complexity parameter, C.

The results, shown in Table 9, indicate that the accuracy of classifying laughter and fussing/crying in children's speech is **77.87%** and **79.37%** respectively. For the tertiary scheme, the accuracy is **69.73%**. These results are significantly better than chance and show that the trained models generalize well to a test set from the MMDB dataset.

Table 9: *Classification results of test set of MMDB using SVM with linear kernel along with the complexity parameter, C, chosen using the grid search.*

Classification Task	Accuracy	Precision	Recall	Complexity (C)
Speech (N=87) vs. Laughter (N=35)	**77.87%**	**74.44%**	**78.51%**	0.059
Speech (N=33) vs. Fussing/Crying (N=30)	**79.37%**	**80.72%**	**85.91%**	2.1
Speech (N=122) vs. Laughter (N=35) vs. Fussing/Crying (N=30)	**69.73%**	**66.15%**	**71.87%**	4.12

The Strange Situation dataset, as mentioned in Section 2.2, has only the fussing/crying events annotated. In order to test the trained models from the MMDB (Speech vs. Fussing/Crying) on this dataset, the features of the speech samples (N = 33) from the MMDB test set were concatenated with the features from the fussing/crying samples (N=62) of the Strange Situation dataset. This can be considered as a cross-corpus testing set. This gives a better sense of the generalization properties of our features and the trained models.

Table 10: *Classification results of using trained models of MMDB (Speech vs. Fussing/Crying) and testing on a cross-corpus test set of MMDB and Strange Situation datasets using SVM with linear kernel and complexity parameter, C=2.1.*

Classification Task	Accuracy	Precision	Recall
Speech(N = 33) vs. Fussing/Crying (N= 62)	**71.6%**	**73.4%**	**71.6%**

The results, shown in Table 10, indicate that trained models generalize well and are capable of discriminating between speech and fussing/crying with an accuracy of **71.6%**. The findings are significant due to the different age groups of the participants, recording conditions, and the type of fussing/crying. The MMDB consists of fussing/crying or whimpering whereas the Strange Situation dataset consists of crying episodes. This indicates that the acoustic features are capable of not only capturing the characteristics of fussing/crying but also that of crying.

7. Conclusions

We have demonstrated the capability of robustly discriminating between children's speech, laughter, and fussing/crying. The combination of wrapper and filter-based features selection algorithms, which encapsulates the intrinsic properties of the dataset and generalizability, has the ability to select acoustic features that are relevant to laughter, fussing/crying, and children's speech. We have shown, through our experiments, that these features have the predictive power to detect laughter and fussing/crying in children's speech. The selected features, trained on samples containing mainly fussing, are capable of robustly detecting crying when tested on to a database with a different age group.

8. Acknowledgments

This work was supported by the National Science Foundation (NSF). (NSF grant No. CCF-1029679)

9. References

[1] W. J. Hudenko, W. Stone, and J.-A. Bachorowski, "Laughter differs in children with autism: an acoustic analysis of laughs produced by children with and without the disorder," *Journal of Autism and Developmental Disorders*, vol. 39, no. 10, pp. 1392–1400, 2009.

[2] G. Esposito and P. Venuti, "Comparative analysis of crying in children with autism, developmental delays, and typical development," *Focus on Autism and Other Developmental Disabilities*, vol. 24, no. 4, pp. 240–247, 2009.

[3] J. Hirschberg, "Dysphonia in infants," *International Journal of Pediatric Otorhinolaryngology*, vol. 49, pp. S293–S296, 1999.

[4] J. Orozco and C. A. R. García, "Detecting pathologies from infant cry applying scaled conjugate gradient neural networks," in *European Symposium on Artificial Neural Networks, Bruges (Belgium)*, 2003, pp. 349–354.

[5] M. K. Rothbart, "Laughter in young children." *Psychological bulletin*, vol. 80, no. 3, p. 247, 1973.

[6] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[7] P. Boersma and D. Weenink, "Praat speech processing software," *Institute of Phonetics Sciences of the University of Amsterdam. http://www. praat. org*.

[8] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "Voicesauce: A program for voice analysis," *Energy*, vol. 1, no. H2, pp. H1–A1.

[9] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Scalaroff, I. Essa, O. Ousley, Y. Li, C. H. Kim, H. Rao, J. Kim, L. Presti, J. Zhang, D. Lantsman, , J. Bidwell, and Z. Ye, "Decoding children's social behavior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*. IEEE, 2013.

[10] M. Ainsworth, M. Blehar, and E. Waters, "Wall. s.(1978)," *Patterns of Attachment: A Psychological Study of the Strange Situation*.

[11] E. Waters, "The reliability and stability of individual differences in infant-mother attachment," *Child Development*, pp. 483–494, 1978.

[12] J. C. Kim, H. Rao, and M. A. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 369–377.

[13] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research*, vol. 37, no. 4, p. 769, 1994.

[14] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[15] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature selection with high-dimensional imbalanced data," in *IEEE International Conference on Data Mining Workshops, 2009. ICDMW'09.* IEEE, 2009, pp. 507–514.

[16] H. Rao, J. C. Kim, A. Rozga, and M. A. Clements, "Detection of laughter in children's speech using spectral and prosodic acoustic features," *Interspeech*, 2013.

[17] H. A. Patil, "Infant identification from their cry," in *Seventh International Conference on Advances in Pattern Recognition, 2009. ICAPR'09.* IEEE, 2009, pp. 107–110.

[18] A. Zabidi, W. Mansor, L. Y. Khuan, R. Sahak, and F. Y. A. Rahman, "Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism," in *5th International Colloquium on Signal Processing & Its Applications, 2009. CSPA 2009.* IEEE, 2009, pp. 204–208.

[19] R. I. Sokol, K. L. Webster, N. S. Thompson, and D. A. Stevens, "Whining as mother-directed speech," *Infant and Child Development*, vol. 14, no. 5, pp. 478–490, 2005.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.